Refresher Course in Calculus, Probability, and Statistics

Day 4: Inferential Statistics

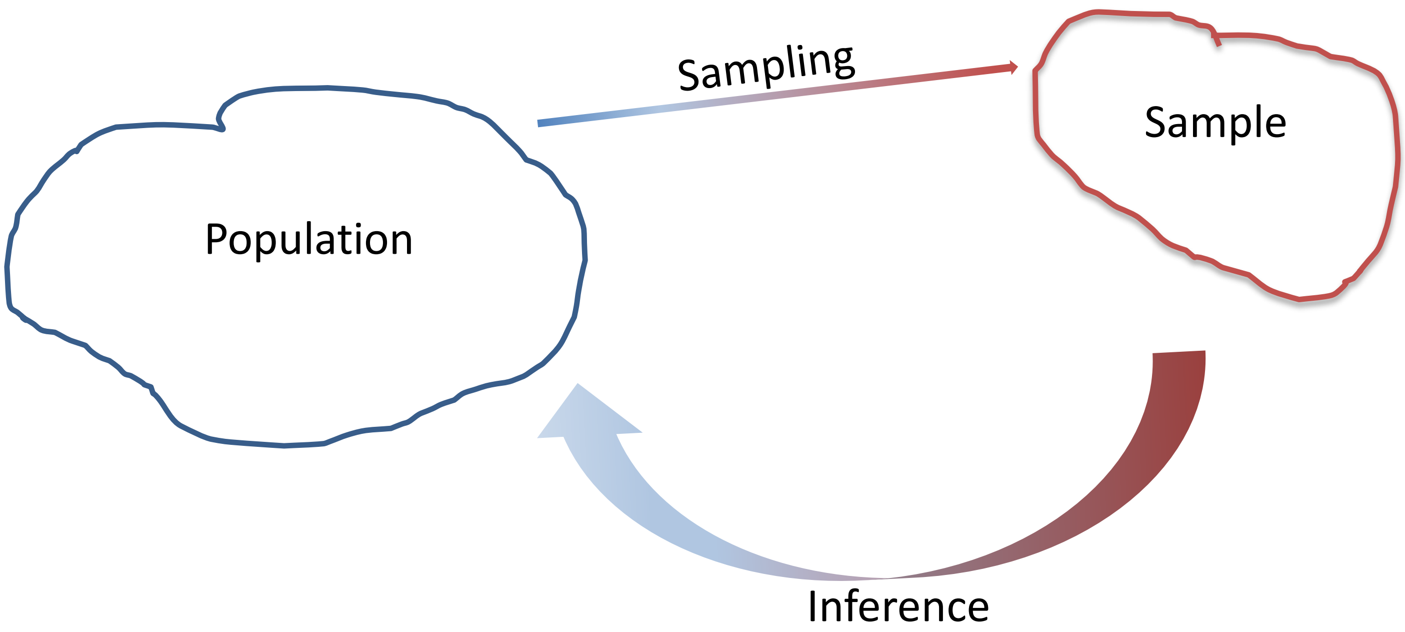
Introduction

"This is when the magic starts happening."
(Marcello Pagano)

- Inferential statistics is the science of using data to learn about the world around us.
- Statistical tools help answer questions about unknown characteristics of distributions in populations of interest:
  - What is the mean earning among recent economics graduates?
  - Do earnings differ for men and women? How much?
- Statistical inference: learn about a population distribution from a random sample.
- Three types of statistical methods:
  - Estimation
  - Confidence Intervals
  - Hypothesis Testing
- Measures of association
  - OLS
- References:
  - [BWA] chap. 7-9; [SWA] chap. 3-7; [HGL] Appendix C

Statistical Inference

- Definition: Use of inductive methods to reason about the distribution of a population on the basis of knowledge obtained in a sample from that population.



Random Sampling

- Samples of a population: Random sampling

- Simple random sampling:
  - Each member of population equally likely to be included in sample
  - Value of random variable Y for i^{th} randomly drawn object denoted Y_i
  - $Y_1, Y_2, \dots, Y_n$: sample of size n from population = data set
  - If the sampling is random, then:
    - Two observations are drawn randomly, thus the value of Y_i contains no information about the value of Y_j: $Y_1, \dots, Y_n$ are *independently distributed*
    - Y_i and Y_j are drawn from the same distribution: $Y_1, \dots, Y_n$ are *identically distributed*
    - → Y_i for $i = 1, \dots, n$ are *independently and identically distributed (i.i.d)*

Random Sampling

Population

Variable	Obs	Mean
age	2,246	39.15316
wage	2,246	7.766949

Sample 1

Variable	Obs	Mean
age	10	39.4
wage	10	6.946883

Sample 2

Variable	Obs	Mean
age	10	39.5
wage	10	8.278549

➲ Many different samples

- many different values of Mean
- A probability distribution of the sample mean

Distribution of the Sample Average

- Sample average: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
- Random sample: $\bar{Y}$ is a random variable
  - Sample average is a random variable and has a probability distribution, called the sampling distribution.
- Suppose observations $Y_1, \ldots Y_n$ are *iid*, and population has mean μ_Y and variance σ_Y^2:
  - Expected average: $E(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} Y_i = \mu_Y$
  - Variance of the sample average: $var(\bar{Y}) = var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}[var(Y_i)] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} cov\left(Y_i, Y_j\right) = \frac{\sigma_Y^2}{n}$
- If $Y_i \sim N(\mu_Y, \sigma_Y^2) \Rightarrow \bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$

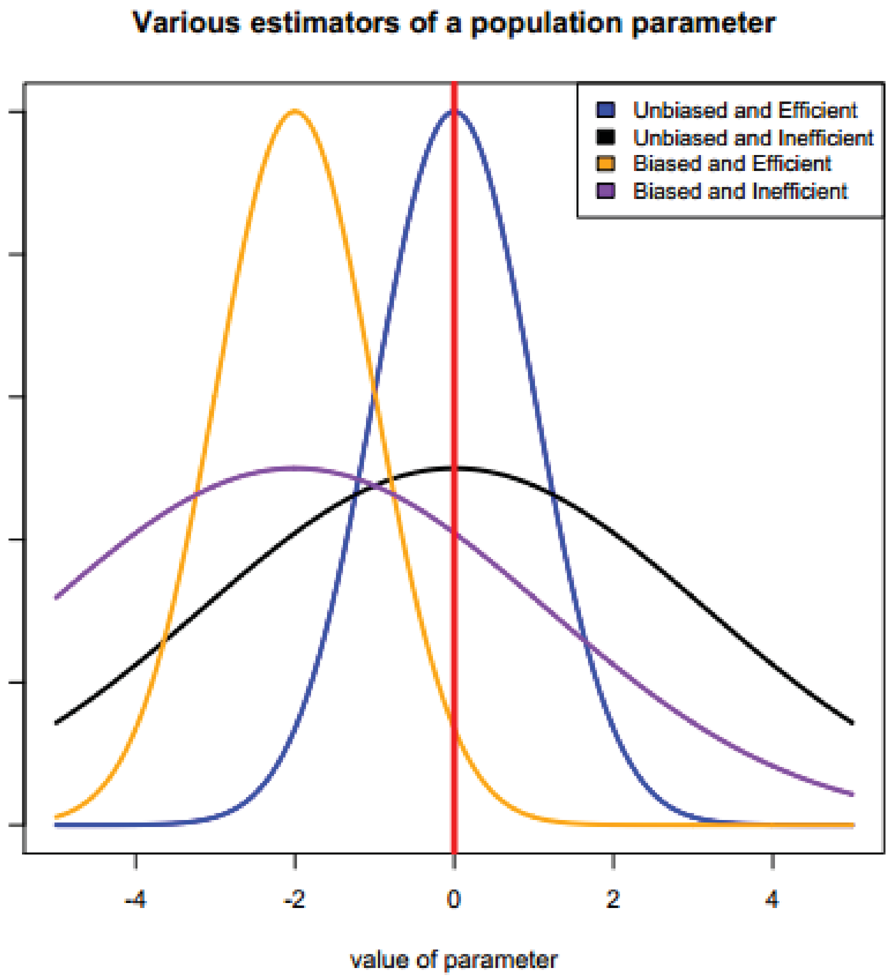
Large Sample Approximations to Sampling Distributions

- **Asymptotic distribution:** large sample approximation to sampling distribution
  - Approximation becomes exact when $n \to \infty$
- **Law of large numbers:**
  - As n becomes large, $\bar{Y}$ will be near μ_Y with a very high probability.
  - **Convergence in probability** or **consistency**: $\bar{Y} \xrightarrow{p} \mu_Y$
- **Central Limit Theorem:**
  - under general conditions, when n is large, distribution of $\bar{Y}$ well approximated by normal distribution
  - Independent of the distribution of Y
  - If Y_i $(i = 1, \ldots, n)$ are iid with $E(Y_i) = \mu_Y$ and $var(Y_i) = \sigma_Y^2$:
  - As $n \to \infty$, $\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$ or $\frac{\bar{Y} - \mu_Y}{\frac{\sigma_Y}{\sqrt{n}}} \sim N(0, 1)$

Estimators and their Properties

- **Estimator**: function of the sample data used to infer the value of a population unknown parameter
  - True (unknown) parameter: μ
  - Estimator: $\hat{\mu}$ (read "mu hat")
- **Estimate**: numerical value of the estimator actually computed from a specific sample
- Desirable characteristics of an estimator:
  - **Unbiasedness**: $E(\hat{\mu}) = \mu$ (Bias: $E(\hat{\mu}) - \mu$)
  - **Consistency**: $\hat{\mu} \xrightarrow{p} \mu$ ($var(\hat{\mu}) \to 0$ as $n \to \infty$)
  - **Efficiency**: between two unbiased estimators $\hat{\mu}$ and $\tilde{\mu}$, $\hat{\mu}$ is more efficient if $var(\hat{\mu}) < var(\tilde{\mu})$

Estimators and their Properties
(continued)



Estimation of Population Mean

- Suppose you want to know the mean value of Y in a population: μ_Y
  - Exactly estimate μ_Y with all values of Y

- You have at hand a sample of n *i.i.d.* observations $Y_1, \ldots, Y_n$ drawn from Y with mean μ_Y and variance σ_Y^2
  - Possible estimators:
    - Y_1
    - $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
    - $\tilde{Y} = \frac{1}{n}\left(\frac{1}{2}Y_1 + \frac{3}{2}Y_2 + \cdots + \frac{1}{2}Y_{n-1} + \frac{3}{2}Y_n\right)$ [assume n is even]

Estimation of Population Mean

How do these estimators fare?

- How do these estimators fare judged by the three criteria?
  - **Unbiasedness**:
    - $Y_1: E(Y_1) = \mu_Y \rightarrow$ unbiased
    - $\bar{Y}: E(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \mu_Y = \mu_Y \rightarrow$ unbiased
    - $\tilde{Y}: \mathrm{E}(\tilde{Y}) = \frac{1}{n}\left(\frac{1}{2}E(Y_1) + \frac{3}{2}E(Y_2) + \cdots + \frac{1}{2}E(Y_{n-1}) + \frac{3}{2}E(Y_n)\right)$

      $= \frac{1}{n}\left(\frac{1}{2}\frac{n}{2}\mu_Y + \frac{3}{2}\frac{n}{2}\mu_Y\right) = \mu_Y \rightarrow$ unbiased

Estimation of Population Mean

How do these estimators fare?

❖ Efficiency:

- $var(Y_1) = \sigma_Y^2 > var(\bar{Y}) = \frac{\sigma_Y^2}{n} \rightarrow \bar{Y}$ is more efficient than Y_1
- $var(\tilde{Y}) = \frac{5}{4}\frac{\sigma_Y^2}{n} > var(\bar{Y}) = \frac{\sigma_Y^2}{n} \rightarrow \bar{Y}$ is more efficient than $\tilde{Y}$
- $var(Y_1) = \sigma_Y^2 > var(\tilde{Y}) = \frac{5}{4}\frac{\sigma_Y^2}{n} \rightarrow \tilde{Y}$ is more efficient than Y_1

❖ $\bar{Y}$ is the most efficient of all unbiased estimators of μ_Y

❖ $\bar{Y}$ is the **B**est **L**inear **U**nbiased **E**stimator (**BLUE**)

- Generally: for $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} a_i Y_i$, where a_i are non-random constants summing to 1, $var(\bar{Y}) < var(\hat{\mu})$

Estimation of Population Variance

- Population variance: $var(Y) = \sigma_Y^2 = E[(Y - \mu_Y)^2]$
  - If we knew μ_Y, we could estimate σ^2:

$$\tilde{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)^2$$

  - But μ_Y unknown → has to be replaced by $\bar{Y}$
  - Unbiased estimator of σ_Y^2:

$$\hat{\sigma}_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

  - Division by $n - 1$ instead of n: degrees of freedom correction (estimating the mean uses one degree of freedom)
- Estimator of population standard deviation σ_Y

$$\hat{\sigma}_Y = \sqrt{\hat{\sigma}_Y^2}$$

How exact are estimates of the population mean?

Population ($\mu_{age} = 39.2\ Years, \mu_{wage} = 7.77\ U.S.Dollars$)

Sample 1

```
 Variable     |         Obs         Mean
-------------+-----------------------
          age |          10         39.4
         wage |          10     6.946883
```

Sample 2

```
 Variable     |         Obs         Mean
-------------+-----------------------
          age |          10         39.5
         wage |          10     8.278549
```

→ Because of random sampling error, impossible to learn exact value of population mean μ.

Standard Error (of the Estimated Mean)

- From CLT: as $n \longrightarrow \infty$, $\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$
- Estimated variance of the estimator $\hat{\mu} = \bar{Y}$:

$$\widehat{var(\bar{Y})} = \hat{\sigma}_{\bar{Y}}^2 = \frac{\hat{\sigma}_Y^2}{n} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n}$$

- Standard Error of $\bar{Y}$:

$$se(\bar{Y}) = \hat{\sigma}_{\bar{Y}} = \frac{\hat{\sigma}_Y}{\sqrt{n}}$$

  - Estimator of standard deviation of sampling distribution of $\bar{Y}$
  - Standard error of the mean
  - Standard error of the estimate

	Mean	Std. Err.
age	39.4	.8969083
wage	6.946883	.9061781

Confidence Intervals for the Population Mean

- **Confidence interval**: range of values that contains μ with a certain pre-specified probability, called confidence level.
  - A 95% confidence interval for μ is an interval constructed so that it contains true value of μ in 95% of all possible random samples.

$$P\left\{\bar{Y} - \left|t^{\alpha}_{(n-1)}\right| \cdot se(\bar{Y}) \leq \mu \leq \bar{Y} + \left|t^{\alpha}_{(n-1)}\right| \cdot se(\bar{Y})\right\} = 1 - \alpha$$

	Mean	Std. Err.	[95% Conf.	Interval]
age	39.4	.8969083	37.37105	41.42895
wage	6.946883	.9061781	4.896966	8.996801

  - Assuming a large sample size (distribution approximately normal)
  - Confidence interval: $P\left\{\bar{Y} - z^{critical} \cdot se(\bar{Y}) \leq \mu \leq \bar{Y} + z^{critical} \cdot se(\bar{Y})\right\}$
    - 90% CI: $z^{critical} = 1.645$
    - 95% CI: $z^{critical} = 1.96$
    - 99% CI: $z^{critical} = 2.58$

Hypothesis Testing

- Proposition about population(s): yes/no question
- Hypotheses we might want to test:
  - Do university graduates earn CHF 6,000 per month on average? More? Less?
    - Hypotheses about the population mean of a single population.
  - Are mean earnings the same for men and women?
    - Hypotheses about differences in means between two populations

- Components of hypothesis testing:
  - Null hypothesis, H_0
  - Alternative hypothesis, H_1
  - Test statistics
  - Rejection region
  - Conclusion

Null and Alternative Hypotheses

- Null hypothesis, denoted H_0 specifies a value c for a parameter:

$$H_0: \quad \mu = c$$

  - H_0 is what we believe until sample provides evidence against it, in which case we reject H_0.

- Alternative hypotheses:

$$H_1: \qquad \mu \neq c \qquad \text{two-sided alternative}$$

$$\left.\begin{array}{ll} H_1: & \mu > c \\ H_1: & \mu < c \end{array}\right\} \quad \text{one-sided alternative}$$

Test statistic

- Sample information about H_0 embodied in a test statistic
- Based on value of test statistic, determine whether it is reasonable to reject H_0 or not
- Consider $H_0: \mu = \mathrm{c}$
  - If sample comes from population $N(\mu, \sigma)$:

$$t = \frac{\bar{Y} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \sim t_{n-1}$$

- When do we use the t distribution and when do we use the normal distribution?
  - In practice, t distribution is only used for small samples.
  - When the sample size is large, the normal distribution is an accurate enough approximation.

Rejection region

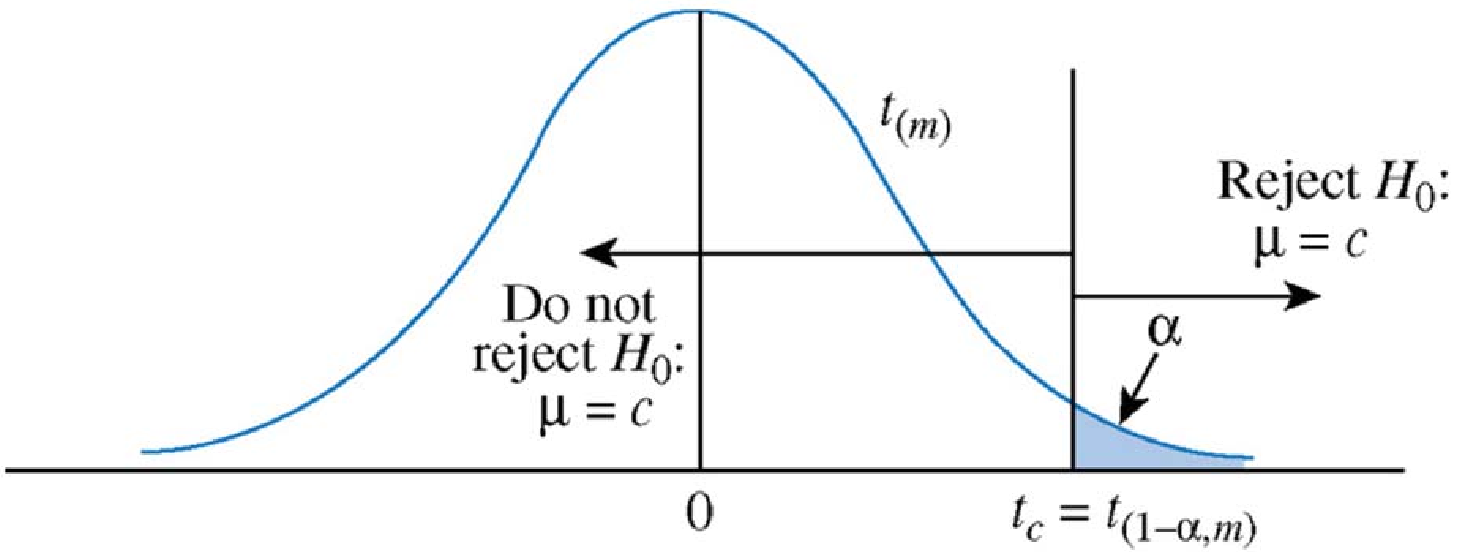
- **Rejection region**: range of values of t-statistic that leads to reject H_0

- If the t-statistic falls in a region of low probability, H_0 is probably not true

- If H_0 is false (H_1 is true), the test statistic is unusually "large" or unusually "small“ given the sample distribution, and depending on the choice of probability α, by which we accept (“fail to reject”) or reject H_0 incorrectly
  - α is called the **level of significance**
  - Common $\alpha's$: 0.1, 0.05, 0.01

One-tailed Test (>)

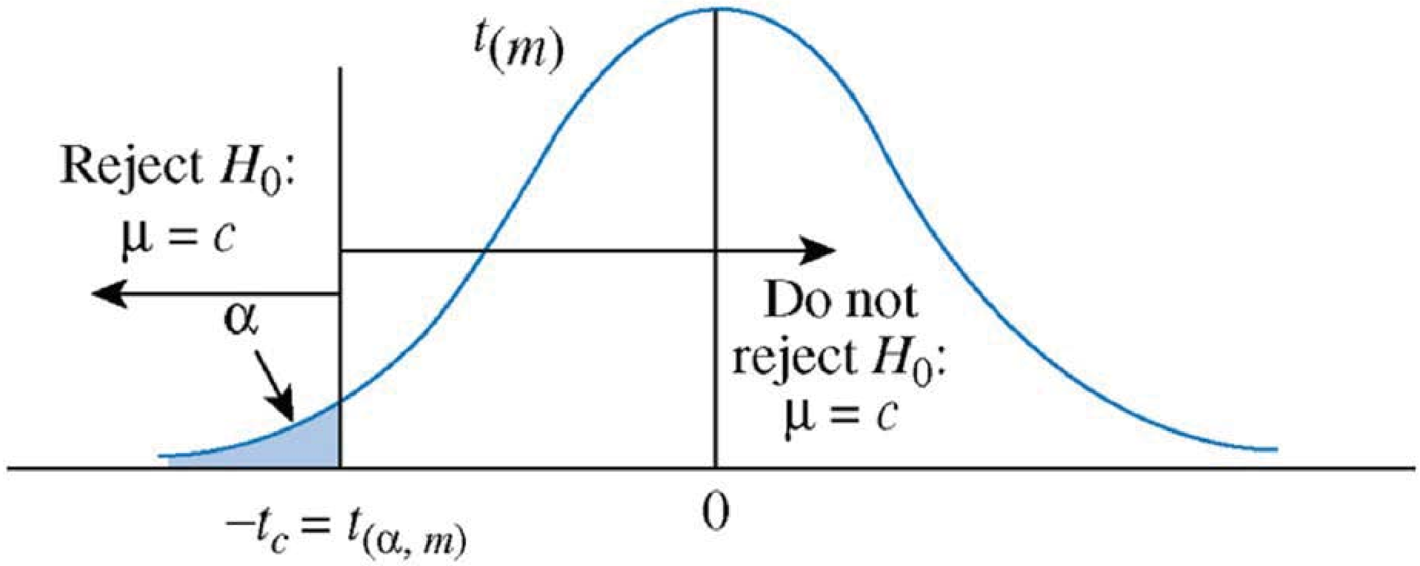
- $H_0: \mu \leq c$ and $H_1: \mu > c$
- Critical value $t_{critical} = t_{(1-\alpha,n-1)}$ is the $100(1-\alpha)$-percentile of a t-distribution with $n-1$ degrees of freedom
  - o $P(t \leq t_{critical}) = 1 - \alpha$
- If the test statistic $\geq t_{critical}$ → reject H_0



Hill, Griffiths, & Lim (2008)

One-tailed Test (<)

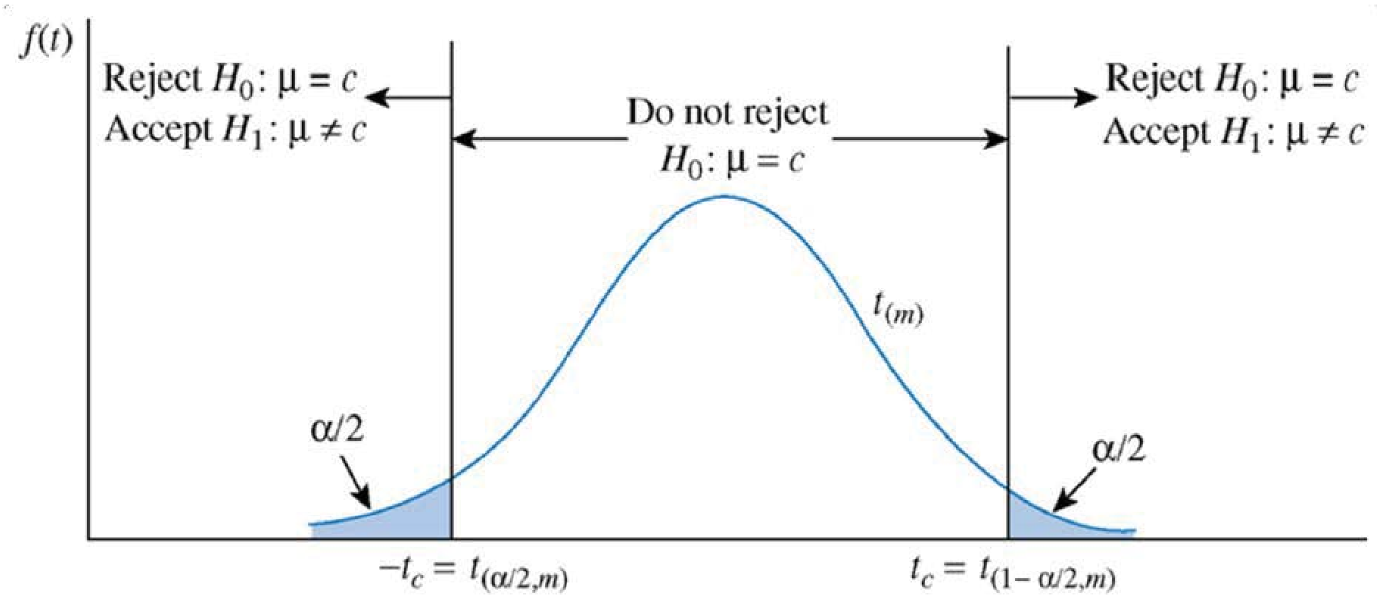
- $H_0: \mu = c$ and $H_1: \mu < c$
- Critical value $-t_{critical} = t_{(\alpha,n-1)}$ is the 100α-percentile of a t-distribution with $n-1$ degrees of freedom
  - o $P(t \leq -t_{critical}) = \alpha$
- If the test statistic $\leq -t_{critical}$ → reject H_0



Hill, Griffiths, & Lim (2008)

Two-tail Test (≠)

- $H_0: \mu = c$ and $H_1: \mu \neq c$
- Critical value $t_{critical} = t_{\left(1-\frac{\alpha}{2}, n-1\right)}$ is the $100\left(1-\frac{\alpha}{2}\right)$-percentile of a t-distribution with $n-1$ degrees of freedom
  - o $P(t \leq t_{critical}) = P(t \geq t_{critical}) = \frac{\alpha}{2}$
- If the test statistic $\geq |t_{critical}|$ or $\leq -|t_{critical}|$ → reject H_0

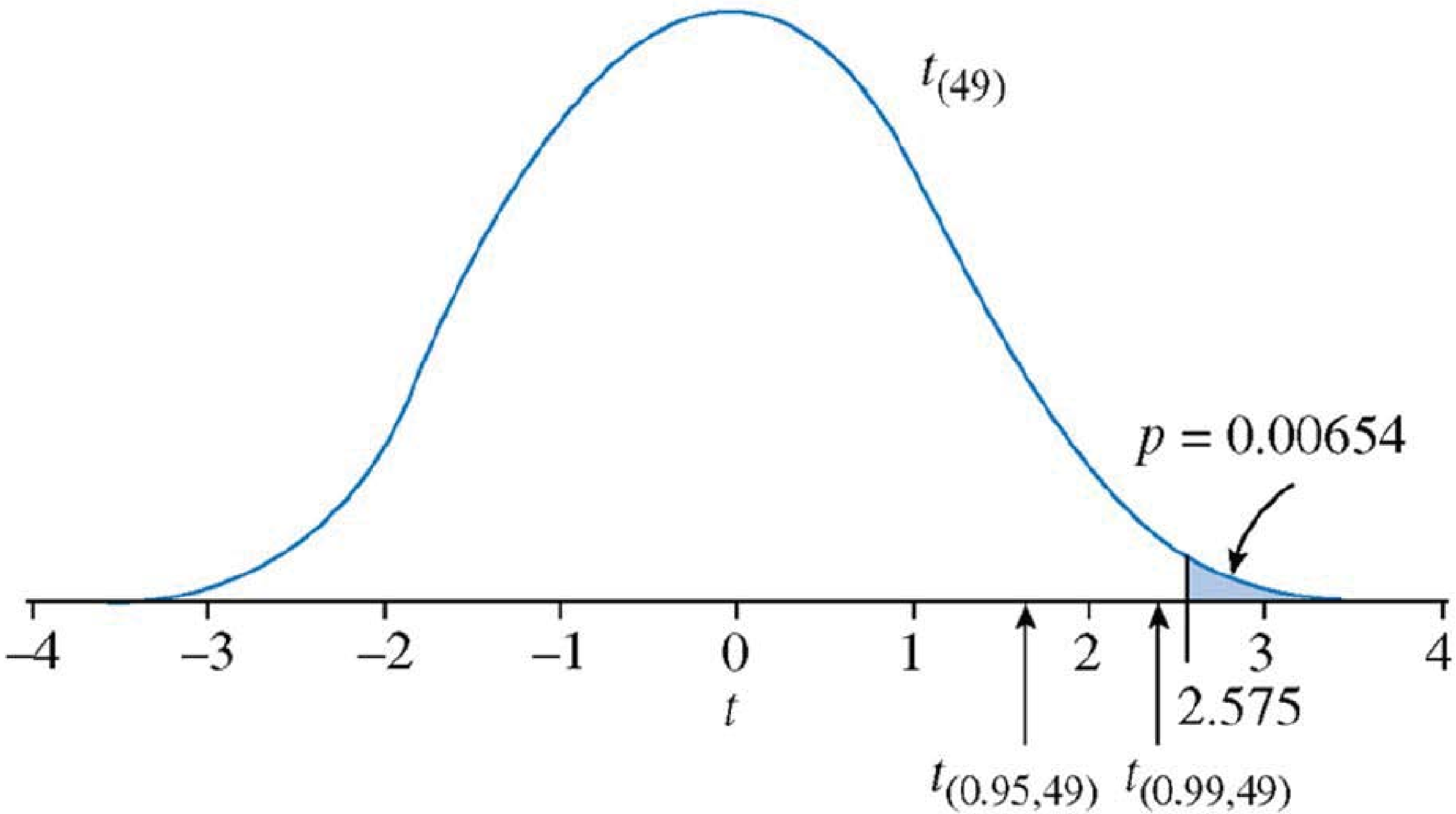


Hill, Griffiths, & Lim (2008)

The *p*-Value

- The *p*-value allows determining outcome of test by comparing it with level of significance α without calculating critical value.
- *p*-value rule:
  - $p < \alpha$: reject H_0
  - $p \geq \alpha$: not reject H_0
- Computation of *p*-value:
  - If $H_1: \mu > c \rightarrow p =$ probability to the right of t
  - If $H_1: \mu < c \rightarrow p =$ probability to the left of t
  - If $H_1: \mu \neq c \rightarrow p =$ sum of probabilities to the right of $|t|$ and to the left of $-|t|$
- When n is large:
  - If $H_1: \mu > c \rightarrow p = 1 - \phi(t)$
  - If $H_1: \mu < c \rightarrow p = \phi(t)$
  - If $H_1: \mu \neq c \rightarrow p = 2\phi(-|t|) = 2[1 - \phi(|t|)]$

p-Value for a One-Tail Test (>)



Hill, Griffiths, & Lim (2008)

Conclusion of the Hypothesis Test

- When hypothesis test completed, conclusion:
  - Reject H_0
  - Fail to reject H_0
  - H_0 is never said to be accepted: absence of evidence is not evidence of absence!
- Do not forget to interpret test results in a meaningful way, given the economic/financial problem you are working on

- Possible mistakes:
  - **Type I error**: H_0 rejected when it is true
  - **Type II error**: H_0 not rejected when it is false

Hypothesis Test concerning the population mean

➲ Is mean income of US women significantly different from $10 per hour?

```
One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    wage |   2,246    7.766949    .1214451    5.755523    7.528793    8.005105
------------------------------------------------------------------------------
    mean = mean(wage)                                             t = -18.3873
Ho: mean = 10                                    degrees of freedom =     2245

    Ha: mean < 10                Ha: mean != 10                Ha: mean > 10
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Testing the Equality of Two Population Means

- Let two populations be distributed as $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$
- To test difference $\mu_1 - \mu_2$, we have to take random samples:
  - Sample of size n_i from population i
  - Sample mean $\bar{Y}_i$ and sample variance $\hat{\sigma}_i^2$
- Null hypothesis: H_0: $\mu_1 - \mu_2 = c$
- Case ❶: Population variances are equal
  - Use both samples to estimate $\sigma_p^2 = \sigma_1^2 = \sigma_2^2$

$$\hat{\sigma}_p^2 = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}$$

  - If H_0 is true:

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - c}{\sqrt{\hat{\sigma}_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)}$$

Testing the Equality of Two Population Means
(continued)

- Case ❷: Population variances are unequal
- If H_0 is true:

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - c}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

  - Exact distribution of this test statistic neither normal nor usual t-distribution
  - Approximated by a t-distribution with degrees of freedom:

$$df = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{\left(\frac{\hat{\sigma}_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{\hat{\sigma}_2^2}{n_2}\right)^2}{n_2 - 1}}$$

  - When both n_1 and n_2 are large, the t-statistic has a standard normal distribution.

Testing the Equality of Two Population Means

(continued)

```
-> white = 0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     ln_wage |        609    1.755669    .5722596    .140951   3.707372

-> white = 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     ln_wage |      1,637    1.910674    .5699247   .0049396   3.693819


Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     609    1.755669    .0231891    .5722596    1.710129     1.80121
       1 |   1,637    1.910674    .0140862    .5699247    1.883046    1.938303
---------+--------------------------------------------------------------------
combined |   2,246    1.868645     .012124      .57458     1.84487     1.89242
---------+--------------------------------------------------------------------
    diff |           -.1550052    .0270814               -.2081126   -.1018979
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -5.7237
Ho: diff = 0                                     degrees of freedom =     2244

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```